Mathematical Theory and Modeling www.iiste.org
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.1, No.1, 2011

Application of Matrix Algebra to Multivariate Data Using Standardize Scores

Aitusi Daniel
Department of Statistics, P.M.B 13, Auchi polytectnic. Auchi.
Phone: +2348032655601 E-mail: aituisidn@yahoo.com

Ehigie Timothy (Corresponding author)
Department of Statistics, P.M.B 13, Auchi polytectnic. Auchi.
Phone: +2348060357105 E-mail: ee4ehigie@yahoo.com

Ayobo Thiophillus.
Department of Statistics, P.M.B 13, Auchi polytectnic. Auchi.
Phone: +2348060357105

Abstract
The aim of this work is to estimate the parameters in a regression equation plane $y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_k X_k + e_i$ *by formulating the correlation matrix R* ($R = X_s' X_s / m$) *and the vector b** ($b^* = R^{-1} r(y)$) *which is the vector of elements between the criterion and each predictor in turn with elements. The parameters were estimated using* $b = b^*(S_y/S_{xi})$ $(i = 1.2)$*. This technique was applied to data extract and the Regression Plane estimated is* $\hat{y} = -2.263 + 1.550x_{1i} - 0.239x_{2i}$ *using the standardized scores.*

Key words: Plane, vector, criterion, correlation matrix, extract, standardized scores.

1. Introduction

We often seek to measure the relationship (if any) between the dependent variable and sets of variables called the independent variables. The data sets collected for the purpose of measurement are usually collected in different units. The use of the original variables measured in the different units were analyzed by (Carrol & Green 1997) using the data on employees absenteeism, attitude towards the time and the number of years employed by the firm. The estimated trend line obtained by them was $\hat{Y} = -2.263 + 1.550x_{i1} - 0.239x_{i2}$. The normal equation formulated by them was in terms of the original data.

(Aitusi & Ehigie 2011) obtained the same trend line using mean-corrected score. (Koutsoyainis 1977) and (Carrol, and Green 1997) both suggested the method of standardized variable which was never applied. In this work, we seek to apply the standardized method to multivariate data, since it is more generalized and can be applied to variables measured in different units.

In multivariate data analysis where we simultaneously estimate the effect of variables on one another, cases of the variables being measured in different units does occur in measuring economic variables. For instance the data set for one variable may be measured in rates while the others in millions, percentages, thousands, hundreds, monetary units etc.

In statistics, a standard score indicates how many standard deviations an observation is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

That is $X_s = (X_i - \overline{X}_i)/S_{xi}$

This process of conversion is called standardizing or normalizing. The standard deviation is the unit of measurement of the z-score. It allows comparison of observations from different normal distribution, which is done frequently in research. Standard scores are also called z-values, z-scores, normal scores and standardized variables. The use of "z" is because the normal distribution is also known as the "z-distribution".

The aim of this work is to estimate the parameters of the regression plane

$$\hat{Y} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e_i \text{ ....(1)}$$

i.e to obtain values for $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ using the standardized scores.

2. Research Methodology

The multiple regression equation for two regressors is $\hat{Y} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e_i$ … (i)

Where e_i satisfies all the required assumptions and (i) is the linear equation for predicting the values of y that minimizes the sum of square errors.

$$\sum_{i=1}^{m} e^2 = \sum_{i=1}^{m} (y - \hat{y})^2 = \text{minimum …(ii)}$$

In this work, we shall estimate the parameters in (i) above as follows;

The correlation matrix R is obtained thus,

$R = X'_s X_s / m$ … (iii)

where X_s is the standardized variable and m is the number of observations.

$b^* = R^{-1} r(y)$ … (iv)

where r(y) is the vector of the product-moment correlations between the criterion and each predictor in turn, with elements.

R^{-1} is the inverse matrix of R

$$r(y) = \frac{Y_s X_{si}}{m} \quad i = 1,2 \quad \ldots \quad (v)$$ Thus, we shall compute

$$r_1 = \frac{Y'_s X_{s1}}{m} \quad \ldots \quad (vi) \qquad r_2 = \frac{Y'_s X_{s2}}{m} \quad \ldots \quad (vii)$$

$$r(y) = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \text{ and } \ldots \quad (viii) \qquad b^* = R^{-1} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \quad \ldots \quad (ix)$$

Hence, $b_1 = b_1^* \left({s_y}/{s_{x1}} \right)$ … (x) $b_2 = b_2^* \left({s_y}/{s_{x2}} \right)$ … (xi)

generally, the parameters are obtained thus $b_i = b_i^* \left({s_y}/{s_{xi}} \right)$, $i = 1,2,...,k \; ...(xii)$

Where s_y and s_{xi} are the standard derivations for variables y and x_i's respectively, and the vector b^* measures the change in y per unit change in each of the predictors when all variables are expressed in standard units.

The equation (xii) will only yield estimates for parameters $b_1 \ldots b_k$ since we are employing the standardized scores, hence, we shall obtain the value of b_0(intercept of the equation) using

$$b_0 = y - b_1 \overline{X}_1 - b_2 \overline{X}_2 - \ldots b_k \overline{X}_k \; \ldots (xiii)$$

3. Data Analysis and Results

Using the same data as (Carrol & Green 1997) we generate data and analyze as in the appendix.

4. Conclusion

The use of standardize scores in data analysis have been greatly emphasized. This is due to the fact that the method converts values in their original form to a new form which is approximately normal. Similarly, the units for which the data collected may be different, hence, the need to standardized the scores to unit-less scores becomes imperative.

Our result is the same as that of (Carrol & Green 1997) using original scores and that of (Aitusi & Ehigie 2011) using mean-corrected score.

References

Aitusi D.N and Ehigie T.O.,(2011) Application of Matrix Algebra To Multivariate Data Using Mean-Corrected Scores. Journal TEMAS Series Vol.2 No.2, kogi State University, Anyigba, Nigeria.

Carroll, D. and Green, J., (1965) "Notes on Factor Analysis" Unpublished Paper, Bell Laboratories, Murray Hill, New Jersey.

J. Douglas Carroll, and Paul E. Green.,(1997); Mathematical Tools For Applied Multivariate Analysis, Academic Press.

Koutsoyiannis A..(1977); Theory Of Economics; An Introductory Exposition Of Econometric Methods, 2nd Edition, New York, palgrave Macmillian

Appendix

5. Data Analysis and Results

	Y	**X_1**	**X_2**	**Y_s**	**X_{s1}**	**X_{s2}**
				$=(y_i - \bar{y})$	$=\dfrac{(X_{1i}-\bar{X}_1)}{S_{X1}}$	$=\dfrac{(X_{2i}-\bar{X}_2)}{S_{X2}}$
	1	1	1	-0.9663	-1.3938	-1.3133
	0	2	1	-1.1503	-1.1283	-1.3133
	1	2	2	-0.9663	-1.1283	-0.9783
	4	3	2	-0.4141	-0.8628	-0.9783
	3	5	4	-0.5982	-0.3319	-0.3082
	2	5	6	-0.7822	-0.3319	0.3618
	5	6	5	-0.2301	-0.0664	0.0268
	6	7	4	-0.0460	0.1991	-0.3082
	9	10	8	0.5061	0.9956	1.0319
	13	11	7	1.2423	1.2611	0.6968
	15	11	9	1.6104	1.2611	1.3669
	16	12	10	1.7945	1.5266	1.7019
TOTAL 75	**75**	**59**				
MEAN 6.25	**6.25**	**4.92**				
STD DEV.	**5.43**	**33.77**	**2.98**			

Note: $Xs_1 = (X_{1i} - \bar{X}_1)/S_{X1}$, $Xs_2 = (X_{2i} - \bar{X}_2)/S_{X2}$

The correlation matrix $R = X'_s X_s/m$

where m = number or paired observation = 12

$$X'_s X_s = \begin{pmatrix} 12.00 & 11.40 \\ 11.407 & 12.00 \end{pmatrix}$$

$$R = (X'_S X_s)/12 = \begin{pmatrix} 12.00 & 11.407 \\ 11.407 & 12.00 \end{pmatrix} \Big/ 12 = \begin{pmatrix} 1 & 0.9506 \\ 0.9506 & 1 \end{pmatrix}$$

$$R^{-1} = \begin{pmatrix} 10.3747 & -9.8620 \\ -9.8620 & 10.3747 \end{pmatrix}, \quad b^* = R^{-1}r(y)$$

where r(y) is a vector $\begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$,thus $r_i = (y_s^1 X_{si})/m$ $r_1 = (y_s^{/} X_{s1})/m$, $r_2 = (y_s^{/} Xs_2)/m$

$$y_s^{/} X_s = \begin{pmatrix} 11.3974 \\ 10.6826 \end{pmatrix} \qquad \therefore r(y) = \begin{pmatrix} 11.3974 \\ 10.6826 \end{pmatrix} /12 = \begin{pmatrix} 0.9498 \\ 0.8902 \end{pmatrix}$$

$$\therefore b^* = \begin{pmatrix} 10.3747 & -9.8620 \\ -9.8620 & 10.3747 \end{pmatrix} \begin{pmatrix} 0.9498 \\ 0.8902 \end{pmatrix} = \begin{pmatrix} 1.0743 \\ -0.1310 \end{pmatrix}$$

$\therefore b^*_1 = 1.0743$ and $b^*_2 = -0.1310$

Thus, the coefficients b_1 & b_2 are obtained as follows

$b = b^* (S_y/S_{xi})$, $S_y = 5.4333$, $S_{x1} = 3.7666$, $S_{x2} = 2.9849$

Hence, $S_y/S_{x1} = \frac{5.4333}{3.7666} = 1.4425$, $S_y/S_{x2} = \frac{5.4333}{2.9849} = 1.8203$

$b_1 = 1.0743 \times 1.4425 = 1.5497 \simeq 1.550$ (3 decimal places)

$b_2 = -0.1310 \times 1.8203 = -0.2385 \simeq -0.239$ (3 decimal places)

we shall obtain b_0 using $b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$

$\therefore b_0 = 6.25 - 1.5497(0.25) - (-0.2385 \times 4.92) = -2.2630$

The estimated regression plane for the multiple regression model is

$\hat{Y} = -2.263 + 1.550x_1 - 0.239x_2$